

Med Zipf mot framtiden En integrerad lexikonresurs för svensk språkteknologi

Lars Borin

Språkbanken
Inst. för svenska språket
Göteborgs universitet

Schæffergårdssymposiet 30/1 2010

bakgrund och förutsättningar

- ▶ lexikonresurser för språkteknologi kräver stora arbetsinsatser för sitt förverkligande
- ▶ resurserna blir användbara om de är interoperabla än om de inte är det, även mellan språk
- ▶ en sorts mångsidigt användbar resurs är ett frasnät (framenet)
- ▶ frasnät finns för några få språk, men inte för svenska
- ▶ Språkbanken har startat ett svenskt frasnätsprojekt, **SweFN++**
- ▶ i det försöker vi återanvända befintliga fria resurser så mycket som möjligt

återanvändning av befintliga resurser

i Språkbanken

- ▶ SALDO (~73.000 betydelser) (**navet i SweFN++**)
- ▶ PAROLE (~29.000 lemgram) (**syntaktisk valens**)
- ▶ SIMPLE (~4000 betydelser) (**semantisk valens**)
- ▶ GLDB/SDB (~60.000 ingångar) (**semantisk valens**)
- ▶ Dalin (1800-t.; ~60.000 ingångar) (**diakroni**)
- ▶ (fornsvenska (~25.000 ingångar) (**diakroni**))
- ▶ (~200 miljoner ord korpusar)

SALDO

Screenshot of the SALDO web application interface. The browser window shows the URL 'http://spraakbanken.gu.se/salido'. The page has a header with the SALDO logo and navigation links. Below the header, there is a section titled 'Introduktion' which contains text about the project's goals and a list of 'Relaterade länkar' (Related links) including 'SveFN++', 'SveFN++', and 'SveFN++'.

[illegible]

The Intercontinental Dictionary Series

Founding Editor:
 Mary Ritchie Key (University of California, Irvine)
 General Editor:
 Bernard Comrie (Max Planck Institute for Evolutionary Anthropology, Leipzig)

Purpose: The purpose of the IDS is to establish a database where lexical material across the continents is organized in such a way that comparisons can be made. Historical studies, comparative, and theoretical linguistic research can be based on this documentation. This is a long term cooperative project that will go on for the next generation or so and will involve linguists all over the world. It is aimed towards interlinear understanding and cooperation. This is a pioneering effort that will have global impact. The purpose also contributes to preserving information on the little-known and "non-prestigious" languages of the world, many of which are becoming extinct.

Rationale: Information on languages of the world is scattered over all the continents and needs to be published in a manner of languages and scripts. There is need of a database where one can find comparable material to formulate hypotheses and test and validate those theories. For example, theories on intercontinental connections have been proposed on the basis of the distribution of word roots and yet there is no single source, where words with this meaning can be found in many languages. Good quantitative and statistical studies are almost impossible to do in a new in-world languages. The IDS will provide a quantitative basis for a scientific approach to language analysis and comparisons. The IDS will provide the research tools necessary for expanding studies such as phonological theory, word formation, language change, lexical distribution, etc.

The screenshot shows a web browser window with the title "The Loanword Typology Meaning List". The address bar displays the URL "http://daniel.scripps.edu/~maphysidat/meanings.htm". The page content is a numbered list of 24 semantic fields, each followed by a link to a corresponding document. The fields are:

- 1 The physical world
- 2 Animals
- 3 AIR/FAIR
- 4 The body
- 5 Food and drink
- 6 Clothing and grooming
- 7 The house
- 8 Agriculture and vegetation
- 9 Basic science and technology
- 10 Motion
- 11 Counting
- 12 Human relations
- 13 Quantity
- 14 Time
- 15 Sense perception
- 16 Emotions and values
- 17 Contrasts
- 18 Speech and language
- 19 Social and political relations
- 20 Warfare and training
- 21 Labor
- 22 Religion and belief
- 23 Random words
- 24 Random lexicon function words

The browser interface includes a search bar, navigation buttons, and a list of open tabs. The page is titled "The Loanword Typology Meaning List" and the content is a numbered list of 24 semantic fields, each followed by a link to a corresponding document.

hopkoppling och harmonisering av lexikonresurser

- ▶ befintliga resurser i flera olika format, med olika sorters innehållskategorier
- ▶ minimimålet är två sorters gemensamma enheter:
 1. betydelser
 2. lemgram (och tillhörande böjningsmönster)
- ▶ vi vill kunna koppla ihop båda sorterna över alla resurser
- ▶ all information måste vara **explicit och entydig**
- ▶ vi vill använda SALDO:s identifierare

stabila identifierare ("PID")

- ▶ i SALDO finns id för:
 - ▶ betydelser (**grad..1**)
 - ▶ lemgram (**grad..nn.1**)
 - ▶ ordklasser (**nn**)
 - ▶ paradigmböjningstabeller (**nn_3u_film**)
- ▶ SALDO-id:na är utformade för att vara:
 - ▶ **unika** (inga andra id behövs, t.ex. som databasnycklar)
 - ▶ **atomära** (inga inbyggda antaganden om betydelser/underbetydelser, etc.) (fast lemgram-id innehåller ordklassbeteckningen)
 - ▶ **användbara i Semantic Web-formalismer** (RDF, OWL): de är välformade xml-namn
 - ▶ **läsbara av människor** (underlättar arbetet med resurserna)

hopkoppling av lexikonresurser

- ▶ harmoniseringen av format kan i stor utsträckning göras med datorprogram
- ▶ hopkopplingen av resursernas innehåll vill vi också automatisera så mycket som möjligt
- ▶ men hur mycket är möjligt? hur mycket manuellt arbete kommer det att innebära?
- ▶ det är här **Zipf** kommer in i bilden

George Kingsley Zipf (1902–1950)



(från Wikipedia)

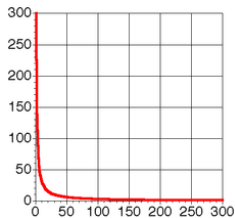


GÖTEBORGS
UNIVERSITET

Språk-
BANKEN

CLT

Zipfs lag: rangordning \Rightarrow frekvens



(från <http://www.useit.com>)

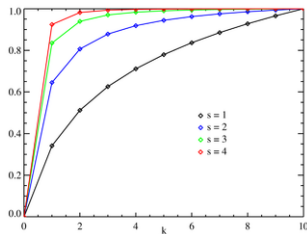


GÖTEBORGS
UNIVERSITET

Språk-
BANKEN

CLT

Zipfs lag: rang \Rightarrow kumulativ mängd



(från Wikipedia)

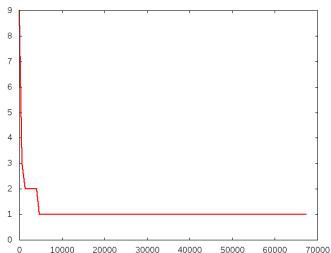


GÖTEBORGS
UNIVERSITET

Språk-
BANKEN

CLT

betydelser per grundform i SALDO

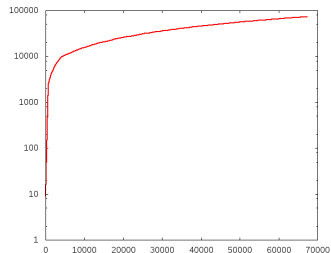


GÖTEBORGS
UNIVERSITET

Språk-
BANKEN

CLT

grundformer \Rightarrow betydelser i SALDO



[illegible]

SynLex – SALDO

max: 92 %	
avg: 80 %	
dev: 0 %	
min: 80 %	
max: 80 %	
avg: 80 %	
dev: 0 %	
min: 80 %	
max: 80 %	
avg: 80 %	
dev: 0 %	
min: 80 %	
max: 80 %	
avg: 85 %	
dev: 6 %	
min: 80 %	
max: 94 %	
avg: 85 %	
dev: 6 %	
min: 80 %	
max: 96 %	
avg: 82 %	
dev: 0 %	
min: 82 %	
max: 82 %	

fräsch.1 fräsch.1

fröjda_sig.1 fröjda_sig.1

fakta.1 fakta.1

stäm.1 tydlig.1 faktastad.1 fullpeppad.1

tänka_etter.1 grunna.1 grubbla.1 betänksa.1 fundera.1 fundera.1 tänksa.1

fundering.1 tänksa.1

SynLex – SALDO, 2

min: 90 %	
avg: 90 %	
dev: 0 %	
min: 94 %	
max: 94 %	
avg: 96 %	
dev: 4 %	
min: 92 %	
max: 100 %	
avg: 94 %	
dev: 2 %	
min: 92 %	
max: 96 %	
avg: 100 %	
dev: 0 %	
min: 100 %	
max: 100 %	
avg: 100 %	
dev: 0 %	
min: 100 %	
max: 100 %	
avg: 90 %	
dev: 0 %	

fröjda.1 fröjda.1

fullpeppad.1 stäm.1

fundamental.1 elementar.1 grundläggande.1

tänksa_etter.1 fundera.1 tänksa.1

kaffa.1 fundera.1 tänksa.1

första.1 biverd.1

hyckelning.1 komb.1

SynLex – SALDO, 3

min: 80 %	
avg: 94 %	
dev: 85 %	
min: 70 %	
max: 100 %	
avg: 82 %	
dev: 0 %	
min: 82 %	
max: 82 %	
avg: 75 %	
dev: 5 %	
min: 70 %	
max: 80 %	
avg: 76 %	
dev: 0 %	
min: 76 %	
max: 76 %	
avg: 80 %	
dev: 0 %	
min: 80 %	
max: 80 %	
avg: 82 %	
dev: 4 %	
min: 80 %	

fräsch.1 tydlig.1 faktastad.1 fullpeppad.1

tänksa_etter.1 grunna.1 grubbla.1 begrunda.1 eftertänksa.1 betänksa.1 reflexion.1 begrunda.1 fundera.1 fundera.1 tänksa.1

fundering.1 tänksa.1

fräsch.1 betänksa.1 fundera.1

fräsch.1 tall.1

fräsch.1 usch.1

fräsch.1 usch.1 tydlig.1 tydlig.1 alkoholist.1 alkoholist.1 alkoholist.1

SynLex – SALDO, 4

avg: 90 %	
dev: 0 %	
min: 90 %	
max: 90 %	
avg: 76 %	
dev: 0 %	
min: 70 %	
max: 86 %	
avg: 80 %	
dev: 0 %	
min: 80 %	
max: 80 %	
avg: 88 %	
dev: 2 %	
min: 86 %	
max: 100 %	
avg: 86 %	
dev: 0 %	
min: 86 %	
max: 86 %	
avg: 80 %	
dev: 0 %	
min: 80 %	
max: 80 %	

fräsch.1 tidning.1

fräsch.1 besvär.1 poesi.1

fräsch.1 hushåll.1

fräsch.1 baracker.1 fräsch.1

fräsch.1 hushåll.1

fräsch.1 spens.1

SweFN-pilotprojekt

- personer: Lars Borin, Dana Dannélls, Markus Forsberg, Maria Toporowska Gronostaj, Annika Kjellandsson, Dimitrios Kokkinakis
- vi har använt existerande standardprogram för snabb start och utprövning av metodologi: Subversion, OpenOffice DB, emacs
- små specialprogram och -skript samt SALDO-webbtjänster (baserade på SALDO/FM-maskineriet) som klister
- återkoppling via automatgenererade webbsidor

SALDO-återkoppling

SALDO Network - Mozilla Firefox

http://www.sprakbanken.gu.se/maskineriet/saldo.html

SALDO

Index: Index de svenska och engelska i SALDO-glossboken.
Så här kommer tillägget till [SALDO](#)

rep.1	hoda.1	PRIM.1	rep.as.1	rep	en	en, en, 1
rep.1	hoda.1	PRIM.1	rep.as.1	rep	en	en, en, 1
rep.2	rep.4	PRIM.1	rep.as.1	rep	en	en, en, 1
rep.5	rep.as.1	PRIM.1	rep.vb.1	rep	en	en, en, 1
enalog.2	digital.1	metastab.1	enalog.av.1	enalog	av	av, 1, g
low.in.rep.1	low.1	PRIM.1	low.in.rep.vb.1	low.in.rep	vib	vib, 1, g
digital.1	odita.1	PRIM.1	digital.av.1	digital	av	av, 1, g
digital.1	dator.1	PRIM.1	digital.av.1	digital	av	av, 1, g
flyg.1	flygplan.1	PRIM.1	flyg.as.1	flyg	en	en, en, 1
flyg.1	flygplan.1	PRIM.1	flyg.as.1	flyg	en	en, en, 1
flyg.2	rosa.1	flygplan.1	flyg.vb.1	flyg	en	en, en, 1
flygplan.1	flygplan.1	PRIM.1	flygplan.en.1	flygplan	en	en, en, 1
flygplan.1	flygplan.1	PRIM.1	flygplan.vb.1	flygplan	en	en, en, 1
hoda.1	hoda.app.1	PRIM.1	hoda.vb.1	hoda	en	en, en, 1
hoda.2	hoda.2	PRIM.1	hoda.vb.1	hoda	en	en, en, 1
hottande.1	hotta.1	PRIM.1	hottande.en.1	hottande	en	en, en, 1
hottande.2	hotta.2	PRIM.1	hottande.en.1	hottande	en	en, en, 1

SALDO-återkoppling, 2

SALDO Network - Mozilla Firefox

http://www.sprakbanken.gu.se/maskineriet/saldo.html

SALDO

Dependency analysis (renewed) - SALDO

No unprocessed parts detected
No additional parts detected
No argument count mismatches detected
No lines of duplicate lines detected
Testbench (2) passed tests, 18 complete tests
23 tests completed

beroendekontroll gentemot SALDO

SveFN++

SveFN++	OK
Dalén	OK
Stenlund	OK
IWT	OK
Parola	OK

- ▶ SALDO-städning (betydelser): nästan klart
- ▶ hopkoppling av SALDO och SynLex: påbörjad
- ▶ detaljerad inventering av fria lexikonresurser: gjord, men inte rapporterad
- ▶ pilotuppsättning frasnätsingångar: klar
- ▶ formatstandardisering (LMF/OWL) och innehållsharmonisering: våren 2010
- ▶ länkning SALDO – Dalin: våren 2010
- ▶ Dalinmorfologi: våren 2010
- ▶ medelsansökan: 10/2 (RJ), 30/3, 20/4 (VR)

metodologi

- ▶ utgå från engelska ramar, modifiera vid behov (samma som SFN)
- ▶ använd de befintliga resurserna så mycket som det går (beroende av god länkning)
- ▶ hitta nya enheter i svenska korpusar

metodologi, 2

- ▶ använd befintliga verktyg för att annotera korpusar så att bra exempelfraser blir lätta att hitta (jfr Deepdict):
 - ▶ MALTParser kan ge (kandidater till) valensramar
 - ▶ SALDO (och annan lexikalisk-semantisk information) kan ge (kandidater till) semantiska grupper av valensargument
- ▶ något som vi är särskilt intresserade av är att utöka mängden flerordningar i våra lexikonresurser

metodologi, 3

- ▶ om arbetet läggs upp på rätt sätt, med rätt datorstöd, så hoppas vi att ny information i lexikonet direkt kan komma korpusökningen till nytta
- ▶ metodologiska och andra forskningsfrågor:
 - ▶ arbetsflöde och verktyg: hur bäst kombinera automatiska processer och manuellt arbete?
 - ▶ semantiska roller och ramelement
 - ▶ ramar och konstruktioner

sammanfattning

- ▶ planerna för SweFN++ omfattar:
 - ▶ återanvändning av fria lexikonresurser för svenska
 - ▶ byggande av ett svenskt frasnät ovanpå resurserna
 - ▶ skapandet av en diakronisk lexikonresurs
 - ▶ utforskande av metodologi för att göra allt detta så arbetsbesparande som möjligt
 - ▶ tillhandahållande av resultatet som en öppen resurs (open content/open source)